INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

On the Price of Anarchy and the Optimal Routing of Parallel non-Observable Queues

Jonatha Anselmi — Bruno Gaujal

N°

Février 2010

ISSN 0249-6399 ISRN INRIA/RR—FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

On the Price of Anarchy and the Optimal Routing of Parallel non-Observable Queues

Jonatha Anselmi , Bruno Gaujal

Thème : Calcul distribué et applications à très haute performance
Équipes-Projets Mescal

Rapport de recherche n° — Février 2010 — 21 pages

Abstract: We consider a network of parallel, non-observable queues and analyze the "price of anarchy", an index measuring the worst-case performance loss of a decentralized system with respect to its centralized counterpart. Our analysis is undertaken from the new point of view where the router has the memory of previous dispatching choices, which significantly complicates the nature of the problem. In the limiting regime where the demands proportionally grow with the network capacity, we provide a tight lower bound on the socially-optimal response time and a tight upper bound on the price of anarchy by means of convex programming. Then, we exploit this result to show, by simulation, that the billiard routing scheme yields a response time which is remarkably close to our lower bound, implying that billiards minimize response time. To study the added-value of non-Bernoulli routers, we introduce the "price of forgetting" and prove that it is bounded from above by two, which is tight in heavy-traffic. Finally, other structural properties are derived numerically for the price of forgetting. These claim that the benefit of having memory in the router is independent of the network size and heterogeneity, while monotonically depending on the network load only. These properties yield simple product-forms well-approximating the socially-optimal response time.

Key-words: Price of anarchy, Optimal routing, Parallel queues, Convex programming, Lambert W function

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Sur le prix de l'anarchie et le routage optimal des files d'attente non-observables et parallèles

Résumé : Dans cet article, nous considérons un réseau de files d'attentes en parallèles, non-observables et nous étudions le prix de l'anarchie, qui mesure la perte de performance pour un système décentralisé par rapport à son analogue centralisé.

Notre analyse introduit un nouveau concept, celui de controleur de routage centralisé avec mémoire, qui change la nature du problème. Dans le régime limite avec des demande qui croissent proportionnellement à la taille du réseau, nous donnons une borne inférieure sur le temps de réponse socialement optimal, ce qui fournit une borne supérieure sur le prix de l'anarchie dans ce contexte.

Ensuite, nous exploitons ce résultat pour montrer par simulation que le routage par suites billards a un comportement remarquablement proche de la borne inférieure. Pour mettre en évidence l'apport de routage non-Bernoulli, nous introduisons le "prix de l'oubli" qui est la rapport entre la performance d'un routage sans mémoire (Bernoulli) et l'un routage avec mémoire, et nous montrons qu'il est borné par 2, cette borne étant exacte dans le cas fortement chargé.

Finalement, d'autres propriétés structurelles sont établies numériquement. Elles mettent en évidence que le bénéfice obtenu grâce à la mémoire du routeur est indépendant de la taille et de l'hétérogénéité du système mais a une dépendance monotone de la charge globale. Cela donne une forme produit simple approximant le prix de l'anarchie de tels systèmes.

Mots-clés : Prix de l'anarchie, Routage Optimal, Files d'attentes parallèles

1 Introduction

The "price of anarchy" [24, 28] is an index measuring the effectiveness of a centralized system with respect to its decentralized counterpart to tradeoff, in service networks, among performance, scalability, and reliability. It is defined as the worst-case response-time ratio between the situation where jobs behave selfishly to maximize their own benefit, yielding the *Nash equilibrium*, and the opposite situation where jobs are controlled by a central authority, yielding the *social optimum* or *social welfare*. While the former identifies the equilibrium point for which any unilateral deviation of each job strategy does not lower its delay, the latter represents the optimal strategy for all jobs in a centralized setting. Therefore, it is the merging of game and queueing theories that provides, in general, the mathematical framework for analyzing this index.

The interest for the price of anarchy in the context of queueing models is currently growing because of its large spectrum of applications: Network routing, load balancing, peer-to-peer and content delivery networks, wireless networks, server farms, grid computing clusters, desktop-grid computing, and database systems; see [29, 27, 22, 20, 13, 34, 18, 9, 1, 4, 6, 30] and the references therein. The great majority of these works provide mathematical tools for characterizing and computing the mean response times in both the situations above and try to relate the price of anarchy to the network size in different settings. This lets designers quantitatively evaluate the loss of performance when shifting to decentralized solutions (corresponding to Nash equilibria) and subsequently perform a suitable dimensioning of the system. In [29] it is shown that the price of anarchy is independent of the network topology as long as the mean jobs arrival rate is less than the mean service rate of the slowest server, and, in this light-load regime, an upper bound is provided. When heterogeneous processor-sharing queues are considered, it is shown in [18, 34] that the price of anarchy, in general, linearly scales with the network size, and it can only depend on the heterogeneity degree of the queues provided that these adopt the shortest-remaining-processing-time scheduling discipline [9]. In the case of multiple central authorities, which can be the case of large server farms, the price of anarchy is shown to be lower bounded by the square root of the number of authorities [4].

We observe that a key point common to all the above works is that the central authority, which in the remainder of the paper we refer to as *router*, achieves the social optimum in a Bernoulli setting, making its routing decisions independent each other, i.e., with no memory. In fact, the social optimum is commonly searched among all the possible Bernoulli policies through a non-linear optimization problem. In several cases, this restriction is known to yield tractable formulas for mean response times, but it may play a non-innocuous role in the analysis. In a more general framework, in fact, the optimal jobs inter-arrival times of each queue are not even i.i.d. [17]. Except for special cases, this *dynamic* routing scheme notoriously complicates the nature of the problem so that the assessment of the routing policy which minimizes the mean response time as well as the analysis of such response time are current open problems in the literature; see, e.g., [17, 5, 10, 14] for the case of parallel and non-observable queues. In this framework, it is also shown in [5] that finding the optimal *cyclic* policy is NP-complete.

1.1 Our Contribution

In this paper, we tackle the problem of analyzing the price of anarchy in open queueing systems of parallel and non-observable queues. We undertake this analysis from a new point of view: In contrast with the existing works above, the key point of our analysis is to consider routers with the memory of previous dispatching choices. This feature of our approach severely complicates the problem of determining the optimal response time achievable by the system, i.e., the social optimum, but we strongly believe that this is necessary for an accurate analysis of nowadays networks. In fact, dispatching schemes with memory, e.g., round-robin, can be easily implemented in network routers with very limited costs. Given the intrinsic intractability of the problem, our analysis is performed in the limiting regime where demands, i.e., jobs arrival rate, proportionally grow with the network size, i.e., the number of queues. This limiting regime is very popular in queueing theory to approximate the behavior of large systems in a tractable manner; see, e.g., [33, 21].

First, we provide a stochastic comparison result providing a tight lower bound on the optimal (mean) response time achievable by the system. This bound is expressed in terms of a convex optimization program which integrates the mean response time of a parallel system of independent $\Gamma/M/1$ queues. New asymptotically-exact bounds on their response times follows as function of the Lambert W function [11].

Then, we introduce the *price of forgetting*, an index measuring the worst-case ratio between the socially-optimal response time of a memoryless router with respect to its memory counterpart. We prove that it is bounded from above by two being exact in heavy-traffic. Thus, the price of anarchy of a router with memory can be understood as the price of anarchy of a Bernoulli router times a correcting factor less than two. When homogeneous queues

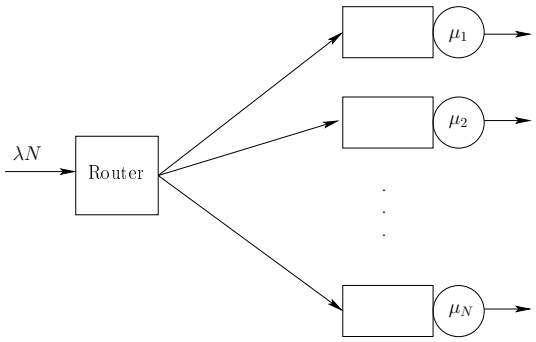

Figure 1: Queueing model under investigation. The router dispatches incoming jobs to the queues according to some policy and with no delay. The mean arrival rate is proportional to the number of queues.

are considered, our analysis simplifies and an explicit expression is found for the price of anarchy which equals the price of forgetting. Here, we prove that it strictly increases in the queues utilization, implying that it is not possible to design a network where the choices of selfish jobs have no impact on performance. This is in contrast with the case of memoryless routers, which make the price of anarchy equal to one for any load [6, 18].

Our analysis also allows us to assess the quality of heuristics for the optimal (non-Bernoulli) routing. An exhaustive numerical analysis reveals that a router forwarding jobs to queues according to a *billiard* scheme, a generalization of round-robin, [3, 16, 19], yields a response time which is remarkably close to our lower bound. In other words, we have the two-fold result that billiard routings achieve, in practice, the minimum response time which, in turn, is very-well captured by our bound and approximations.

Finally, we give numerical evidence of the fact that the price of forgetting is insensitive to the network size (N) and heterogeneity, while monotonically depending on the network load (L) only. These structural properties entail that the price of anarchy admits the product-form $f(N)F(L)$ where i) $f(N)$ is linear and refers to the price of anarchy achieved by a Bernoulli router (which is well-understood [6, 18]), and ii) $F(L)$, the price of forgetting, is increasing in L and bounded from above by two. The term $F(L)$, explicitely given, is thus interpreted as the added-value of a router with memory or, more technically, as the response-time improvement which is obtained when the input process of each queue is Gamma instead of Poisson.

This paper is organized as follows. Section 2 introduces the model under investigation and the necessary preliminaries. In Section 3, we propose an upper bound on the price of anarchy in terms of a convex program as well as an improved approximation. In Section 4, we define the price of forgetting which is analyzed to derive qualitative properties on the benefit of a router with memory. In Section 5, we show how a non-Bernoulli router should operate to minimize response time, and, in Section 6, we measure its impact on system performance exhibiting structural properties that we interpret. Finally, Section 7 draws the conclusions of this work.

2 Model and Preliminaries

We consider a queueing system composed of N infinite-room queues working in parallel as shown in Figure 2. Jobs arrive from an external Poissonian source having intensity λN to a router which instantaneously dispatches jobs to one of the N queues according to a given *policy*, i.e., routing rule.

We assume that the router cannot observe the state of the queues, i.e., their current number of jobs.

In queue i, we assume that jobs require service for an exponentially-distributed amount of time having mean $\mu_i^{-1} = O(1)$. The service times are i.i.d. and independent of the arrival process and N. Initially, all queues are supposed to be empty (this assumption can be relaxed because the mean performance does not depend on initial states, but it is useful in our proofs for technical reasons). The scheduling discipline of each queue is assumed to be First-Come-First-Served or Processor Sharing. In the remainder of the paper, index i implicitly ranges from 1 to N, if not otherwise specified.

The routing policy $a \stackrel{\text{def}}{=} (a^1, \ldots, a^N)$ of jobs into queues is given by the sequences $(a_n^i)_{n\in\mathbb{N}} \in \{0,1\}$, where $a_n^i = 1$ if the n^{th} job is sent to queue i, and is 0 otherwise. By definition, if $a_n^i = 1$ then $a_n^j = 0$ for all $j \neq i$, i.e., a job is routed to a single queue.

In contrast with existing works, we analyze the case where the router has memory of which queues previous jobs have been dispatched. This feature of our model is innovative in the context of evaluating the price of anarchy, easy to implement in real-world routers and plays a key role in our analysis.

Let

$$L \stackrel{\text{def}}{=} \lambda N / \sum_{i=1}^{N} \mu_i \tag{1}$$

be the *network load*, an index measuring the "network utilization". Within the optimal routing policy, the considered queueing model is stable if and only if $L < 1$.

We also denote by $R = R(a)$ the mean response time, or sojourn time, of jobs in the system under policy a, provided the expectation exists (the dependence of a will be reported when necessary). In the remainder of the paper, we omit the words "mean" when we refer to response time for simplicity.

2.1 Nash Equilibrium and Social Optimization

Within the model introduced above, we consider two different scenarios. In the first one, selfish jobs choose to join a queue to minimize their response time individually, and we refer to this situation as *Nash equilibrium*; see, e.g., [6, 18]. The response time achievable in this scenario is denoted by R^{Ne}. In the second one, jobs choose to join a queue to minimize the response time of all jobs taking into account choices of previous jobs. We refer to this situation as *social optimization*, and the mean response time achievable in this scenario is denoted by R^{So}. These scenarios reflect the conflicting situations where jobs moves in an infrastructure with neither control nor shared information with respect to the case where a centralized object controls the dynamics of the system to maximize the profit of all jobs.

Our notion of social optimization differs from the one considered in existing approaches in the sense that we let the router operate with the memory of previous decisions. This means that the set of policies handled by the router is much larger than the set of the Bernoulli ones, because the routing decisions are no more independent each other. The main consequence of this point is that the optimal jobs inter-arrivals of each queue are not i.i.d. [17], and, thus, the analysis becomes much more difficult. In the case of Nash equilibrium, however, we observe that the optimal policy must be Bernoulli because jobs make their decisions independently of the others (in fact, no shared information is available in a fully-decentralized system before the arrivals of jobs). As a consequence, existing works apply to our model in this case (see [6] for a formula for R^{Ne}).

It is clear that both the situations depicted above can be modeled in our queueing system by specifying a suitable policy in the router.

We measure the efficiency of the social optimization scenario with respect to the Nash equilibrium by means of the price of anarchy A, which is defined as follows

$$A \stackrel{\text{def}}{=} R^{Ne}/R^{So}. \tag{2}$$

Evidently, large values of A indicate that the impact of a centralized control drastically improves the performance of the system, and vice versa. On the other hand, we notice that a centralized control design is less scalable and reliable than a distributed one because the system has a single point of failure. We also stress that A represents the *worst-case* delay inefficiency of an uncontrolled infrastructure with respect to a controlled one. By definition, $A \geq 1$.

In contrast with the analysis of R^{Ne}, there are currently no exact analyses for the efficient computation of R^{So} when N is generic (see the introduction). Even though R^{So} could be determined, in our settings, by applying Markov decision process algorithms, e.g., [26], the intrinsic computational complexity of this approach makes it feasible only when the number of queues is very limited. In the particular case where $N = 2$, an efficient analysis of R^{So} can be found in [14]. This current difficulty prevents the understanding of the problem and motivated the authors to investigate alternative computational techniques for the approximate analysis of (2).

3 Analysis

In this section, we develop an approximation and a lower bound on the socially-optimal response time by means of convex programming. Approximations and bounds on the price of anarchy immediately follow by (2).

3.1 Bounding the Optimal Response Time

We now introduce a lower bound on the socially optimum response time R^{So}.

Let $q_n^i(a_1^i, \ldots, a_n^i)$ be the amount of work in queue i after n arrivals. We denote by

$$Q_n^i(a_1^i, \ldots, a_n^i) \stackrel{\text{def}}{=} \mathbb{E} q_n^i(a_1^i, \ldots, a_n^i), \tag{3}$$

the mean work where the expectation is taken over all arrival times and service times. Now, let $R_i(a^i)$ be the Cesaro limit of Q_n^i, i.e.,

$$R_i(a^i) = \lim_{m\to\infty} \frac{1}{m} \sum_{n=1}^{m} Q_n^i(a_1^i, \ldots, a_n^i). \tag{4}$$

Using the PASTA property, e.g., [7], $R_i(a_i)$ is the mean response time of the jobs sent to queue i.

For any $0 < \delta < 1$, let $p_\delta^i \stackrel{\text{def}}{=} (1-\delta)\sum_{k=1}^{\infty} \delta^{k-1} a_k^i$, that exists since all a_n^i are bounded. By definition of p_δ^i, $\sum_{i=1}^N p_\delta^i = 1$. Therefore, the set $\mathcal{L}$ of limit points of $(p_\delta^1, \ldots, p_\delta^N)$, when $\delta \to 1$, also has a sum equal to one.

We also define the regular sequence with rate p and phase θ: for all $n \geq 1$, $\alpha_n^i(p,\theta) \stackrel{\text{def}}{=} \lfloor np+\theta \rfloor - \lfloor (n-1)p+\theta \rfloor$. Note that $\alpha_n^i(p,\theta) \in \{0,1\}$ for all n as long as $p \leq 1$ and is periodic in θ with period 1.

Theorem 1 *Under the foregoing notations, the mean response time of a job under policy a verifies*

$$R(a) \geq \inf_{(p_1,\ldots,p_N)\in\mathcal{L}} \big(p_1 R_1(\alpha(p_1,0)) + \cdots + p_N R_N(\alpha(p_N,0))\big).$$

The theorem says that the average response time of any policy is larger than the combination of response times in all queues where the arrival process in each queue is a regular sequence. This result is to be compared with [16] where regular sequences with rate r are proved to be optimal admission sequences in a single queue under the constraint that a proportion of at least r packets have to be admitted in the queue. The main difference comes from the fact that routing to several queues is more difficult than admitting to a single queue because one does not know whether the proportion of jobs sent to each queue by the optimal routing policy exists. This is still an open problem and Theorem 1 above does not answer to this question but just provides a lower bound on the response time of the optimal policy. For the proposed lower bound, such proportions exist in all queues and are equal to the rates p_i achieving the infimum. On the other hand, the result stated in 1 is very close to Theorem 25 in [2]. The main difference is the fact that our cost is not additive, making the proof slightly more involved.

Let us consider a single queue i and the arrival process induced by $\alpha(p_i,\theta)$ in queue i. Let $k \stackrel{\text{def}}{=} \lfloor 1/p_i \rfloor$. The inter-arrival process $\tau_1, \cdots, \tau_n, \cdots$ has a distribution that is made of a sequence of sums of k (or $k+1$) i.i.d. exponential distributions, with parameter $N\lambda$. For example, if $p_i = 2/7$, then $k = 3$ and the arrival process in queue i under policy $\alpha(2/7, 0) = 0,0,1,0,0,0,1,0,0,1,\cdots$ where the sequences $0,0,1$ and $0,0,0,1$ alternate, is such that the distribution of the inter-arrival times alternates between the sum of three exponentials with rate $N\lambda$ and the sum of four exponentials with rate $N\lambda$.

In general, the arrival rate in queue i is $p_i N\lambda$. Now , considering a stationary i.i.d. arrival process $T_1, \cdots T_n \cdots$, with a Gamma distribution for inter-arrival times, with parameters p_i and $N\lambda$. It should be clear that for any n, these two inter-arrival processes compare for the convex ordering of random sequences: $(\tau_1, \cdots, \tau_n) \geq_{cx} (T_1, \cdots, T_n)$.

Using the fact that the mean response time of jobs is a convex increasing function of the input process, this implies that the mean response time in queue i under a regular arrival process with rate p_i is larger that the mean response time in queue i under a Gamma-distributed arrival process with rate $p_i N\lambda$. Putting this together with Theorem 1 yields the following result.

Theorem 2 *Let $R_i^{\Gamma(a,b)/M/1}$ be the mean response time of a job in queue i having exponential service times and i.i.d. inter-arrival times with a $\Gamma(a,b)$ distribution. Then,*

$$R^{So} \geq \inf_{\substack{\pi_1,\ldots,\pi_N \geq 0: \\ \pi_1+\cdots+\pi_N=1}} \sum_{i=1}^{N} \pi_i R_i^{\Gamma(1/\pi_i, N\lambda)/M/1}. \tag{5}$$

In the following, we will use a coefficient that scales with N for the proportion of jobs sent to queue i: we define $\beta_i \stackrel{\text{def}}{=} N\pi_i$, where β_i is a positive constant. A lower bound on R^{So} is finally obtained by solving the following

optimization problem

$$GB(N) \stackrel{\text{def}}{=} \begin{array}{ll} \min & \sum_{i=1}^{N} \frac{\beta_i}{N} R_i^{\Gamma(N/\beta_i, N\lambda)/M/1}(\beta_i) \\ \text{s.t.} & \\ & \sum_{i=1}^{N} \beta_i = N \\ & U_i(\beta_i) \leq 1, \ \forall i \\ & \beta_i \geq 0, \ \forall i, \end{array} \tag{6}$$

where

$$U_i(\beta_i) = \lambda \beta_i / \mu_i \tag{7}$$

and $GB(N)$ stands for Gamma-Bound with N queues. By means of Little's law, the quantity U_i is interpreted as the *utilization* of station i, and it represents the "proportion of time" in which station i is busy (in the long term), e.g., [25].

3.1.1 Heavy-Traffic Behavior

The proposed bound $GB(N)$ is interpreted as the response time achieved when the input arrival processes of all queues are independent and Gamma distributed. This means that the router can be now thought as Bernoulli, provided that its jobs arrival process is no more Poisson. Given that all queues become independent $\Gamma/M/1$ queues, classic heavy-traffic analysis immediately applies to derive useful approximation and insights; see, e.g., [23, 15].

This corollary follows by Theorem 2 and the heavy-traffic analysis of GI/GI/1 queues; e.g., [23].

Corollary 1 *As $L \to 1$, $GB(N) \geq R^{So}_{Bernoulli}(N)/2$, where equality holds when N is large.*

In other words, our bound is essentially half of the response time of the optimal Bernoulli routing in heavy-traffic. This reveals one important structural property that we anticipate here and analyze in next sections for the non-heavy-traffic case: The added-value of a router with memory is independent of the network heterogeneity and size.

3.2 Asymptotic Analysis of the $\Gamma/M/1$ Queue

In previous section, we established a lower bound on the social optimum R^{So} in terms of an optimization problem involving the response time of parallel $\Gamma/M/1$ queues. The integration of the exact $\Gamma/M/1$ (or $G/M/1$) analysis, see [7], in the constraints of (6) renders a non-linear problem which seems to be difficult to analyze, e.g., in terms of convexity, and also yields numerical instabilities related to $O(N^N)$ terms. Therefore, we now address the development of simple (non-numerical) approximations for the response time of $\Gamma/M/1$ queues, which, to the best of our knowledge, are not currently available in the literature (except for the heavy-traffic case).

The following theorem provides bounds on $R_i^{\Gamma(N/\beta_i, N\lambda)/M/1}$ for any network load and size.

Theorem 3 *Let $\sigma_i, {\sigma_i}^+ \in [0, 1)$, respectively, be the (unique) solutions of the equations*

$$z exp(\tfrac{1-z}{U_i}) = 1 \tag{8}$$

and

$$z exp(\tfrac{1-z}{U_i}) \left(1 - \frac{1}{2} \frac{(1-z)^2}{N U_i^2}\right) = 1. \tag{9}$$

Then,

$$\frac{1}{\mu_i(1-\sigma_i)} \leq R_i^{\Gamma(N/\beta_i, N\lambda)/M/1} \leq \frac{1}{\mu_i(1-\sigma_i^+)}. \tag{10}$$

Given that, as $N \to \infty$, $\sigma^+ \to \sigma$, the following corollary is straightforward.

Corollary 2 *As $N \to \infty$,*

$$R_i^{\Gamma(N/\beta_i, N\lambda)/M/1}(\beta_i) \to \frac{1}{\mu_i(1-\sigma_i)} \tag{11}$$

from above.

Lemma 1 $\sigma_i \leq U_i^2$.

Rewriting (8) as

$$-\frac{z}{U_i}e^{-\frac{z}{U_i}} = -\frac{1}{U_i}e^{-\frac{1}{U_i}} \tag{12}$$

and observing that it admits exactly two positive roots when $0 \leq U_i \leq 1$, where the largest one is at $z = 1$, we note that σ_i can be expressed in terms of the Lambert W function [11] if and only if $-z/U_i \geq -1 = -W(-1/e)$, which is true by Lemma 1. Hence,

$$\sigma_i = -U_i W(-\frac{1}{U_i}e^{-\frac{1}{U_i}}). \tag{13}$$

where W is the principal Lambert function (with $W(0) = 0$).

We recall that the Lambert W function [11], defined as the inverse function of $f(W) = W\exp(W)$, over $[-1, +\infty)$, satisfies

$$0 \leq -W\left(-\frac{1}{U_i}e^{-\frac{1}{U_i}}\right) \leq 1 \tag{14}$$

for all $0 \leq U_i \leq 1$. In particular, $-W(-1/e) = 1$ and $W(0) = 0$. Therefore, by means of Formula (11) and within the considered limiting regime, the response time of a queue with Gamma-distributed arrivals and exponential service times can be interpreted as the response time of an exponential queue with Poisson arrivals and utilization multiplied by the correcting factor $-W(-\exp(-1/U_i)/U_i)$, which plays a very important role in our analysis of the price of anarchy. In the particular case where $N/\beta_i \in \mathbb{N}$, we observe that the popular $E_{N/\beta_i}/M/1$ queue is found.

The rate of convergence of Formula (11) is strictly related to the convergence of $(1 + a/N)^N$, for a fixed, to its limiting value $\exp(a)$ (see the proof of Theorem 3), which is known to be $\Theta(1/N)$. In the experimental results section, we numerically show that this suffices to obtain very accurate response time estimates even when N is relatively small and that it provides improved accuracy with respect to heavy-traffic approximations.

We also observe that the result of Theorem 2 can be extended to higher-order moments of response time. This is ensured by the uniform convergence of the states stationary probabilities which follows by the analysis above.

As last remark, we note that the applicability of Formula (11) goes beyond the scope of this paper and immediately finds applications in capacity planning studies, e.g., optimization of power consumption in server farms with utilization and/or quality of service (QoS) constraints, and in the analysis of (bulk service) $M/M^{N/\beta_i}/1$ queueing systems because a strict relation with $E_{N/\beta_i}/M/1$ queues is known, e.g., [7].

3.3 Approximations for Large Network Sizes

The simplicity of Formula (11) allows for the development of a simple optimization procedure. In fact, problem (6) can be rewritten as follows

$$\begin{aligned} GB(\infty) \stackrel{\text{def}}{=} \quad & \min \quad \frac{1}{\lambda N}\sum_{i=1}^{N}\frac{U_i}{1-\sigma_i(U_i)} \\ & \text{s.t.} \\ & \quad \textstyle\sum_{i=1}^{N}\frac{\mu_i}{\lambda}U_i = N \\ & \quad 0 \leq U_i \leq 1, \ \forall i, \end{aligned} \tag{15}$$

where $\sigma_i(U_i)$ is given by (13), which is exact as $N \to \infty$.

Remark 1 *For any N, $GB(\infty) \leq GB(N) \leq R^{So}(N)$.*

Let also $GB_N(\infty)$ be the optimum of (15) where $\sigma_i(U_i)$ is given by (9). Even though $GB_N(\infty)$, in general, does not seem to provide upper bounds on the optimal response time, the following result ensures that it always provides improved accuracy with respect to $GB(\infty)$ when estimating R^{So}.

Theorem 4 $R^{So} - GB(\infty) > |R^{So} - GB_N(\infty)|$.

Even though more accurate approximations than $GB(\infty)$ and $GB_N(\infty)$ for R^{So} can be derived (by taking into account more expansions terms, see the proof of Theorem 4), we will numerically show that they suffice to obtain very accurate results.

The following result ensures that efficient algorithms, see [8], can be immediately applied to solve (15) in polynomial time.

Theorem 5 *The optimization problem* (15) *is convex.*

The proposed upper bound on the price of anarchy (2) simply follows by taking the ratio between R^{Ne} [6] and $GB(\infty)$. A numerical evaluation of its tightness and convergence speed is postponed in the experimental results section.

4 Price of Forgetting

We define the "price of forgetting" as the worst-case ratio between the socially-optimal response time achieved with a Bernoulli router and a router with memory, i.e.,

$$F \stackrel{\text{def}}{=} R^{So}_{Bernoulli}/R^{So}. \tag{16}$$

The following connection is immediate.

Proposition 1 *The price of anarchy* (2) *is given by the product between the price of forgetting* (16) *and the price of anarchy achieved with a memoryless router.*

Since the price of anarchy in the Bernoulli case is well-understood, we limit the focus on the price of forgetting.

4.1 Heterogeneous Queues

The nature of the price of forgetting in heavy-load conditions immedately follows by the discussion in Section 3.1.1. Coincisely, for any network size

$$\lim_{L\to 1} F(L) \leq 2. \tag{17}$$

In contrast, in light-load conditions we must have

$$\lim_{L\to 0} F(L) = 1, \tag{18}$$

which is intuitive. In fact, if the queue lengths are almost empty, then the response time approaches the service time of the fastest queue in any case.

The following result extends the result (17) over all network loads.

Theorem 6 *For any network load and size,* $F \leq 2$.

These results imply that the socially-optimal Bernoulli policy yields a response time which can be at most twice larger than the response time achieved by a router with memory. We will numerically show that F is an increasing function of the network load.

4.2 Homogeneous Queues

A scenario of practical interest is the case where the queues are homogeneous, i.e., $\mu_1 = \cdots = \mu_N = \mu$, for which we can draw additional results and easily compare with the case of Bernoulli routers.

Remark 2 *If the router has no memory, then the socially optimum response time coincides with the response time in Nash equilibrium (see [18, 12]) and they both admit a very simple formula, i.e.,*

$$R^{So}_{Bernoulli} = R^{Ne} = \frac{1}{\mu(1-L)}, \tag{19}$$

where $L = U = \lambda/\mu$.

In other words, the price of anarchy, in the context of memoryless routers, becomes one regardless of the utilizations.

Remark 3 *If the router has memory, then* (19) *implies that the price of forgetting equals the price of anarchy (when the queues are homogeneous).*

We now apply our analysis in this easier setting in order to understand the price of forgetting (or price of anarchy). The following result is known in the literature (and also follows from Theorem 1 using a symmetry argument); see, e.g., [32] Prop. 8.3.4.

Theorem 7 ([32]) *Under the foregoing assumptions, the round-robin (or cyclic) policy minimizes the mean response time for any N.*

The results which follow in the remainder of this section are implicitly assumed to hold in the considered limiting regime, i.e., when $N \to \infty$.

The following result is an immediate consequence of Theorems 3, 7 and Formula (19), and provides an asymptotically-exact formula for the price of anarchy.

Corollary 3

$$F(L) = A(L) = \frac{1 + LW(g(L))}{1 - L} \tag{20}$$

where $g(L) = -exp(-1/L)/L$ and $L = U = \lambda/\mu$.

As first consequence, we notice that the price of anarchy now becomes a function of the utilization U (note that L boils down to U here), which is in contrast with the case of memoryless routers. In other words, *it is not possible to design a network where the behavior of selfish jobs has no impact on the response time as in the memoryless case*. Formula (20) is thus interpreted as the correcting factor that should be taken into account by a Bernoulli analysis of the price of anarchy. Secondly, the expression (20) allows us to derive more results than Theorem 6.

Corollary 4 *$A(L)$ is strictly increasing in L and*

$$\lim_{L\to 0} \frac{dA(L)}{dL} = 1, \quad \lim_{L\to 1} \frac{dA(L)}{dL} = 0. \tag{21}$$

The limits in (21) show that i) the response-time benefits of a router with memory are non-negligible even when the utilizations are small, and that ii) $A(L)$ is concave in heavy-load conditions (we will numerically show that concavity does not seem to hold for $A(L)$ in general), and, thus, large improvements can be obtained even in a non-negligible neighborhood of $L = 1$.

5 Optimal Routing

The framework introduced above allows us to numerically inspect the response-time gap between our bounds and approximations and heuristic policies for the optimal routing.

In this section, we first perform a validation of Formula (11) on several models. Then, we measure the performance achieved with a router assigning jobs to queues according to billiard sequences, e.g., [3, 16, 19]. We show that the resulting distance from our formulas is remarkably small. Coincisely, our conclusions are that i) *the billiard routing scheme minimizes the response time*, and ii) *our bounds and approximations on the minimum response time are tight.*

5.1 Accuracy of Formula (11)

We now measure the accuracy of asymptotic formula (11) by means of the percentage relative error

$$\frac{|R^{\Gamma/M/1}_{exact} - R^{\Gamma/M/1}_{approx}|}{R^{\Gamma/M/1}_{exact}} 100\%, \tag{22}$$

where $R^{\Gamma/M/1}_{exact}$ is obtained numerically through the (exact) standard analysis of the $G/M/1$ queue, and $R^{\Gamma/M/1}_{approx}$ is given by (11). Numerical computations have been performed using Maple 13. We initially evaluate (22) by varying $N \in \{50, 100, 200, 1000\}$ and $U \in \{0.1, 0.2, \ldots, 0.9, 0.95\}$. Since the mean arrival rate λ affects the percentage relative error (22) only through the utilization, it is not considered in our experiments.

Figure 2 illustrates the quality of (22) in the above cases. As N grows, we first note that the accuracy of (11) increases, which is expected because it is asymptotically exact. For $N = 50$, (11) is remarkably accurate and yields a relative error always less than 2%. Since real-world systems are composed of hundreds (or thousands) of queues, we conclude that the asymptotic formula (11) can be used to obtain very accurate response time estimates of the $\Gamma/M/1$ queue.

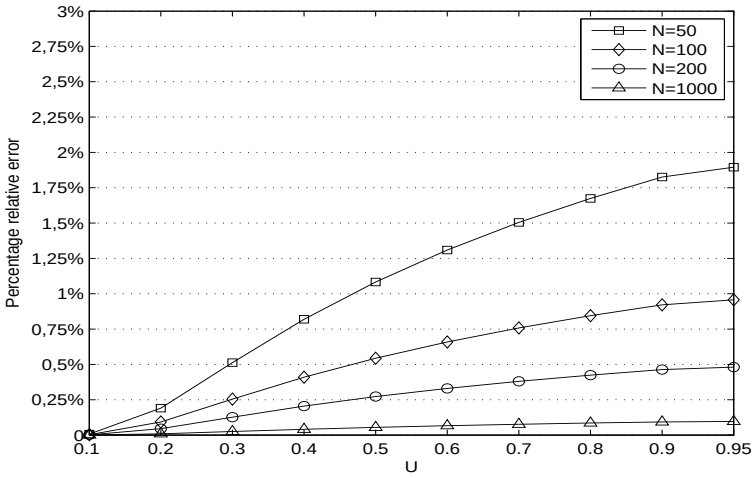

Figure 2: Accuracy evaluation of the asymptotic formula (11) through the error measure (22).

5.2 Quasi-Optimality of Billiard Sequences

We consider the case where the router forwards jobs to queues according to billiard sequences, see [3, 16, 19], which are constructed through the SG algorithm introduced in [19] (easily implementable in network routers with a very limited cost). The SG algorithm takes has input the fraction of jobs to send to the queues (given by the solution of (15)) and an initial-position vector $x \in \mathbb{R}^N$ which we assume such that $x_i = 1$ if $\mu_i = \max_j \mu_j$ and 0 otherwise (we point the reader to [19] for further details on the SG algorithm and billiard sequences). Given that a numerical solution of the response time induced by billiard sequences is impractical for a number of reasons, e.g., the aperiodicity of the resulting routing patterns, we use simulation. To measure the gap between the response time achieved with this routing scheme and our bounds/approximations, we assess the general quality of the percentage relative error

$$\mathrm{Err}_{\mathrm{App}} = \frac{|R_{\mathrm{App}} - R_{\mathrm{Sim}}|}{GB_N(\infty)} \cdot 100\% \tag{23}$$

where $R_{\mathrm{App}} \in \{GB(\infty), GB_N(\infty)\}$ (defined in Section 3.3) axnd R_{Sim} is the average response time computed by simulation. We measure percentage relative errors with respect to $GB_N(\infty)$ because it represents the closest approximation of R^{Opt} (see Theorem 4). The measures of R_{Sim} refer to 99% confidence intervals having size no larger than 1% of R_{Sim} itself. For any pair (N, L), $N \in \{20, 50, 100\}$ and $L \in \{0.10, 0.15, 0.20, \ldots, 0.95\}$, we generated 1,000 random models where the service rates μ_i have been drawn in the range $[0.01, 100]$ according to a uniform distribution. Larger values of N have not been considered because of the strong computational requirements of simulation. In any case, the proposed analysis sufficies to assess the accuracy of our approach.

The experimental results of this analysis are summarized in Figures 5.2, which, thus, refers to a total of nearly 50,000 experiments. In the figure, the dashed (continuous) lines refer to the error obtained with $GB(\infty)$ ($GB_N(\infty)$) for different network sizes. We clearly see that the response time achieved through a billiard routing is remarkably close to our approximation $GB_N(\infty)$ and also to our bound $GB(\infty)$. Given that the optimal response time achievable by the system must lie between our bound and the response time achieved by the billiard routing, we conclude, in an empirical sense, that billiard sequences are optimal for the response time which is very-well approximated by our analysis.

5.3 Computational Requirements

We illustrate the computational requirements for calculating the price of anarchy through the mathematical program (15). These are important to know because the program (15) should be executed each time the network changes to reinitialize the parameters of the optimal routing algorithm. Real-world systems change over time for a number of reasons, e.g., addition or removal of one queue, variation of the arrival rate or service times, CPU frequency scaling.

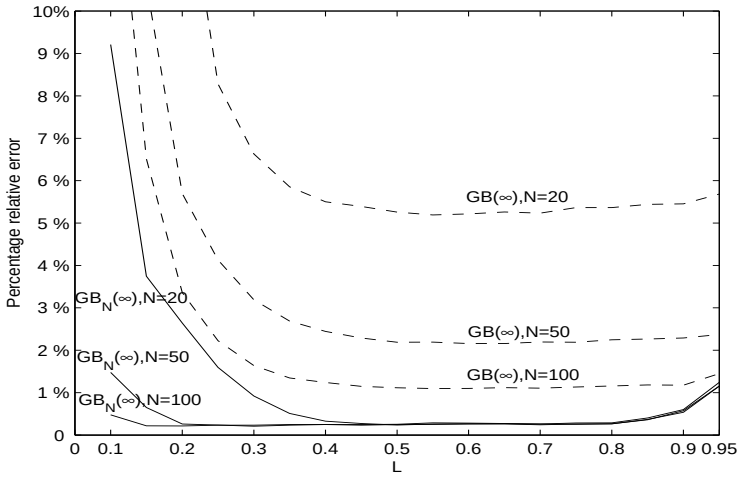

Figure 3: Plots of the error (23) averaged over a large number of tests.

Experiments have been performed by running the Ipopt optimization solver [31] on a 2.80Ghz Intel Xeon processor with multi-threading technology. By varying N, we consider a wide test-bed of randomly generated models, where the service rates uniformly range in $[0.01, 100]$ and the arrival rate is such that the network load (1) uniformly ranges in $[0, 1]$. Table 1 illustrates the average and the standard deviation of the computation times (in seconds) required by the solution of (15), where each number refers to a sample of 1,000 models. From the

N	50	100	500	1,000	5,000
Average time (sec)	0.181	0.251	1.317	2.433	10.85
Std. dev.	0.038	0.059	0.187	0.312	2.901

Table 1: Computation times required by the solution of (15) by varying N.

results in the table, we conclude that the solution of (15) is almost online: models of networks composed of a thousand of servers only require, in the average, less than three seconds.

6 The Impact of Routers with Memory: Structural Properties

We now measure the proposed upper bound on the price of forgetting in order to numerically investigate its fundamental properties. Following the results of previous section, it is very tight. We infer an important structural property which we interpret: *the price of forgetting only depends on the network load L, meaning that it is independent of the network heterogeneity and size.*

6.1 Homogeneous Queues

In the case of homogeneous queues, the proposed bound boils down to the simple formula (20), which is asymptotically exact. By varying the utilization from 0.05 to 0.95 with step 0.05, Figure 4 illustrates i) the asymptotic price of anarchy (20) (the dashed bold line), ii) the price of anarchy obtained with a memoryless router (the dashed-dotted line), and iii) for $N \in \{10, 50, 100, 1000\}$, the exact price of anarchy, which is obtained by applied standard analysis of the $E_N/M/1$ queue. In that figure, we first notice that the price of anarchy is not concave and (slightly) increases as N does converging to our asymptotic formula (20). The fact that the price of anarchy increases with N finds the simple intuition that adding new resources gives more and more freedom to the router for optimizing the response time with respect to its Bernoulli counterpart. On the other hand, the price of anarchy in the case of memoryless router remains constant to one for any number of queues and utilizations, and it is remarkably far from the ones where the router has memory. The exact price of anarchy computed for $N = 100$ is very close to our asymptotic formula and, for $N = 10$, it has almost the same behavior. When $N = 100$

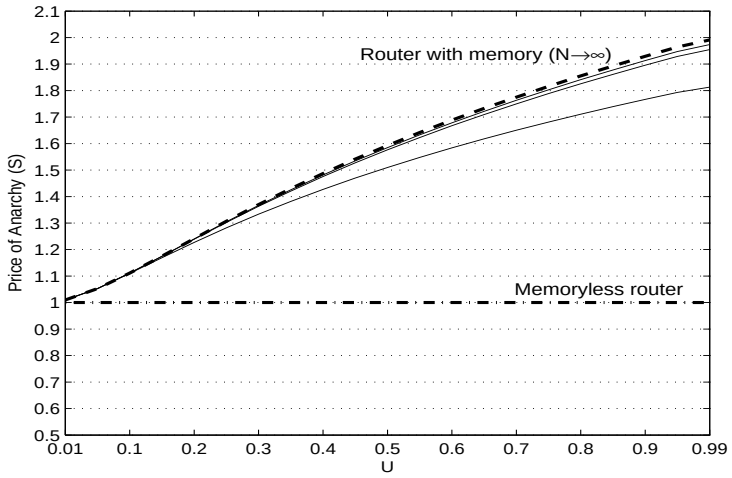

Figure 4: Price of anarchy (20) (equivalent to price of forgetting here) by varying the queues utilization. The three continuous lines correspond to the exact prices of anarchy for increasing network sizes, where the lowest refers to $N = 10$ and the largest to $N = 1000$.

and $U = 0.85$, Figure 4 shows that a Bernoulli-based analysis underestimates the price of anarchy of a factor 1.9. When the utilizations are 0.1, i.e., small, the Bernoulli price of anarchy is 10% lower. These observations immediately quantify how large can be the worst-case impact of considering routers with memory in the design of distributed or centralized systems, where utilizations usually range in $[0.6, 0.85]$.

6.2 Heterogeneous Queues: Independence of Network Heterogeneity and Size

We now measure the price of forgetting in the heterogeneous case. We first consider an illustrative example which we use to inspect fundamental properties. Then, we carry out an extensive numerical analysis to give evidence of their correctness.

An illustrative scenario We consider a clustered network composed of N queues where 1/10 of the queues have fast service rates $\mu_f = 100$, 2/10 of the queues have medium service rates $\mu_m = 50$, and the remaining ones have low service rates $\mu_l = 1$. By varying the network load (L) and size (N), we plot the resulting price of forgettings in Figure 5, which lets us draw two important hypotheses.

First, we observe that *our bound on the price of forgetting is independent of the network size* (note that this is also true for the case of homogeneous queues, provided that N is sufficiently large, see Figure 5).

Second, if the ratios of Figure 5 are compared pointwisely to the corresponding ones of Figure 4 (where the concepts of network load and utilization are equivalent) we note that these points are very close each other. This suggests that *our bound on the price of forgetting is not influenced by the heterogeneity of the considered scenario*, as L varies, becoming a function of the network load only. In Section 3.1.1, we showed that this property holds true in heavy-traffic and as $N \to \infty$.

Exhaustive numerical investigation We now carry out an extensive numerical analysis to give evidence of the independence of the price of forgetting on the network size and heterogeneity. To do this, we focus on a very large test-bed of randomly generated models drawing the service rates μ_i in $[0.01, 100]$ uniformly. For any pair (N, L), we generated 1,000 models computing average and standard deviation of the price of forgetting. The results of this analysis are shown in Table 2, which refers to a total of 48,000 different models. The results presented in that table robustly confirm the two hypotheses arisen in previous section. When $N = 50$, we observe that the averages of the price of forgetting are already settled to their asymptotic value. Furthermore, the standard deviations are very small and decreasing in both N and L. This shows the independence with respect to the network heterogeneity. By varying L and U (for $L = U$), Figure 6 plots (20) and the average price of forgetting shown in Table 2 to stress independence with respect to heterogeneity. Both curves are remarkably close each other, and they are almost equivalent when $L \geq 0.55$. In the figure, we observe that the slight gap achieved when

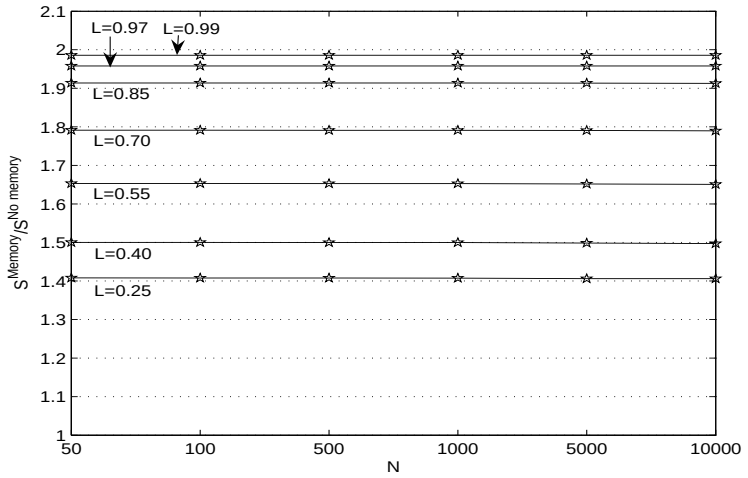

Figure 5: Insensitivity of the price of forgetting with respect to network size.

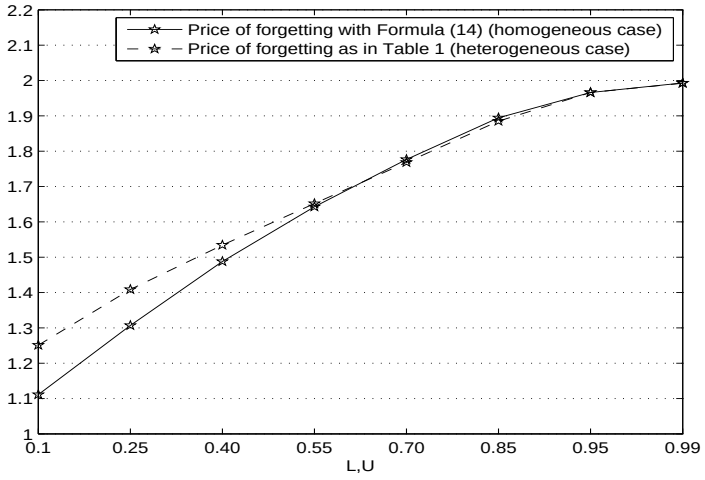

Figure 6: Comparison of Formula (20) with the averages of the prices of forgetting in Table 2.

Averages

N	50	100	500	1,000	5,000	10,000
$L = 0.10$	1.252	1.254	1.254	1.254	1.253	1.253
$L = 0.25$	1.408	1.409	1.409	1.409	1.409	1.409
$L = 0.40$	1.534	1.534	1.535	1.534	1.535	1.534
$L = 0.55$	1.652	1.652	1.652	1.652	1.652	1.652
$L = 0.70$	1.768	1.768	1.768	1.768	1.768	1.768
$L = 0.85$	1.885	1.885	1.885	1.885	1.885	1.885
$L = 0.95$	1.966	1.966	1.966	1.966	1.966	1.966
$L = 0.99$	1.992	1.992	1.992	1.992	1.992	1.992

Standard deviations

N	50	100	500	1,000	5,000	10,000
L =0.10	3.0e-2	2.0e-2	8.3e-3	7.9e-3	6.9e-3	6.1e-3
L =0.25	1.4e-2	1.0e-2	5.3e-3	4.8e-3	2.8e-3	1.8e-3
L =0.40	9.5e-3	6.9e-3	3.1e-3	2.3e-3	2.1e-3	8.9e-4
L =0.55	5.3e-3	3.9e-3	1.7e-3	1.3e-3	7.1e-4	6.4e-4
L =0.70	2.9e-3	2.1e-3	1.0e-3	8.3e-4	5.5e-4	5.3e-4
L =0.85	2.1e-3	1.6e-3	7.7e-4	6.4e-4	4.8e-4	4.5e-4
L =0.95	7.3e-4	5.9e-4	4.7e-4	4.5e-4	4.4e-4	4.3e-4
L =0.99	4.8e-4	4.6e-4	4.3e-4	4.4e-4	4.3e-4	4.2e-4

Table 2: Averages and standard deviations of our bound on the price of forgetting over the large number of tests (e-n reads 10^{-n}).

L is small must go to zero as $L \to 0$ because, in this regime, the optimal Bernoulli and non-Bernoulli response times equal the service time of the fastest queue. Also, the fact that the curve is increasing follows by Theorem 6.

Except for the heavy-traffic case in Section 3, this is surprising because *the optimal fractions of jobs sent to each queue in the Bernoulli and non-Bernoulli settings are different* (see next section). These structural properties show that:

- our bound on the price of anarchy can be seen as the product between the price of anarchy with a memoryless router and (20). Equivalently,
- our bound (15) on the optimal response time can be seen as the ratio between the optimal Bernoulli response time and $F(L)$ given by (20).

The tightness of our GB bounds provides the following approximation for the optimal response time:

$$R^{So} = R^{So}_{Bernoulli}/F(L). \tag{24}$$

6.3 Optimal Routing Probabilities Comparison

We show the relation between the routing probabilities of the optimal Bernoulli router (p_i) and of our bound (15) (π_i) by evaluating the quantity $\sum_{i=1}^{N} |\pi_i - p_i|$, which measures their cumulative absolute difference, over the experiments performed in previous section. While in heavy-traffic the fractions of jobs in a memory/non-memory setting are equal (which is obvious) and the properties above could find some interpretation, this does not hold for the non-heavy-traffic case, for which a significant difference exists (see Table 3). Notwithstanding, the price of forgetting is not affected by such difference as shown in previous section.

L	0.25	0.40	0.55	0.70	0.85	0.95	0.99
	1.9e-1	7.7e-2	2.6e-2	6.2e-3	5.4e-4	1.3e-5	$\leq$e-5

Table 3: $\sum_{i=1}^{N} |\pi_i - p_i|$ by varying the network load (e-n reads 10^{-n}).

7 Concluding Remarks

We presented a new framework for assessing the performance benefits of large centralized infrastructures with respect to their decentralized counterparts through the price of anarchy. Our analysis lets the central router exploit local information on its past routing decisions to achieve the social optimum. We showed that the price of anarchy admits can be interpreted as the product between the corresponding memoryless price of anarchy and an increasing function of the network load only approaching two in heavy-traffic. The latter represents the added-value of having memory in the router. Also, we used our framework to compare routing policies for the optimal response time, numerically showing that the response time achieved by a billiard routing scheme, which provides an upper bound, matches our lower bound. We leave as future work the case with general service time distributions.

8 Proofs of Our Results

Proof of Theorem 1

First, note that Q_n^i is only defined on integer points in $\{0,1\}^n$ and can be extended to $[0,1]^n$ by linear interpolation over simplexes, defined by the multimodular base $v_k = (0,\dots,0,-1,+1,0,\dots,0)$ (see [17]). Once this is done, it has been shown that Q_n^i has the following properties [2]:

(*i*) Q_n^i is convex.

(*ii*) for all m, $Q_n^i(a_1^i,\dots,a_n^i) = Q_{n+m}^i(0,\dots,0,a_1^i,\dots,a_n^i)$.

(*iii*) for all $m<n$, $Q_n^i(a_1^i,\dots,a_n^i) \geq Q_m^i(a_{n-m+1}^i,\dots,a_n^i)$.

The last two properties are easy to verify since they are also true on each trajectory, for the quantities ℓ_n^i: Point (*ii*) is true because the system is initially empty and time-homogeneous, while the third property comes from the fact that adding a job in the past increases the load at time 0.

However, the first item (*i*) is not true for the random quantities q_n^i, but only in expectation.

Using these properties, one has:

$$
\begin{aligned}
&\sum_{n=1}^{\infty}(1-\delta)\delta^n a_n^i Q_n^i(a_1^i,\dots,a_n^i)\\
&\geq \sum_{n=1}^{M}(1-\delta)\delta^n a_n^i Q_M^i(0,\dots,0,a_1^i,\dots,a_n^i)\\
&+\sum_{n=M+1}^{\infty}(1-\delta)\delta^n a_n^i Q_M^i(a_{n-M+1}^i,\dots,a_n^i)\\
&\geq \Big(\sum_{n=1}^{\infty}(1-\delta)\delta^n a_n^i\Big) Q_M^i\Big(\sum_{n=1}^{M}(1-\delta)\delta^n(0,\dots,0,a_1^i,\dots,a_n^i)\\
&+\sum_{n=M+1}^{\infty}(1-\delta)\delta^n(a_{n-M+1}^i,\dots,a_n^i)\Big)\\
&= p_\delta^i Q_M^i(\delta^M p_\delta^i, \delta^{M-1}p_\delta^i,\dots,p_\delta^i).
\end{aligned}
$$

.

Now, let us define

$$B_i(M,\delta,p) \stackrel{\text{def}}{=} \sup_{0\leq\delta\leq 1} Q_M^i(\delta^M p,\delta^{M-1}p,\dots,p), \tag{25}$$

and let us consider all the queues together. By definition,

$$R(a) = \limsup_{K\to\infty}\frac{1}{K}\sum_{n=1}^{K}(a_n^1 Q_n^1(a^1)+\dots+a_n^N Q_n^N(a^N)). \tag{26}$$

Using the well-known fact that the Cesaro limit is always larger than the discounted limit with a discount going to one,

$$\begin{aligned} R(a) &\geq \limsup_{\delta\to 1}(1-\delta)\sum_{n=1}^{\infty}\delta^n\left(a_n^1 Q_n^1(a^1)+\cdots+a_n^N Q_n^N(a^N)\right) \\ &\geq \limsup_{\delta\to 1}\sum_{i=1}^{N} p_\delta^i Q_M^i(\delta^M p_\delta^i, \delta^{M-1}p_\delta^i,\ldots,p_\delta^i) \\ &\geq \inf_{(p_1,\ldots p_N)\in\mathcal{L}}\sum_{i=1}^{N} p_i Q_M^i(p^i,p^i,\ldots,p^i). \end{aligned}$$

The point $(p^i,p^i,\ldots,p^i)$ belongs to the simplex of $\mathbb{R}^M$ whose extreme points are all the regular sequences of length M: $\alpha(p_i,\theta)$, $0\leq\theta\leq 1$. Since Q_M^i is linear on each simplex, one gets $Q_M^i(p^i,p^i,\ldots,p^i)=\mathbb{E}_\theta Q_M^i(\alpha(p_i,\theta)_1,\ldots,\alpha(p_i,\theta)_M)$, where the expectation is taken over the uniform distribution on $0\leq\theta\leq 1$.

Finally, by letting M go to infinity, it has been shown in [2] that one gets, for the Cesaro limit,

$$\begin{aligned} &\lim_{M\to\infty} Q_M^i(p^i,p^i,\ldots,p^i) \\ &\geq \lim_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m} Q_n^i(\alpha(p_i,0)_1,\ldots,\alpha(p_i,0)_n). \end{aligned}$$

This concludes the proof.

Proof of Corollary 1

In heavy-traffic,

$$R_i^{\Gamma(N/\beta_i,N\lambda)/M/1}\approx\lambda N\frac{1/(\beta_i\lambda^2 N)+1/\mu_i^2}{2(1-U_i)}+\frac{1}{\mu_i}, \tag{27}$$

where the summands in the numerator of the first term refer to the variance of interarrival and service times. After some algebra, $GB(N)$ has thus the form (for some β_i)

$$GB(N)=\sum_{i=1}^{N}\frac{\beta_i}{N}\frac{2-U_i+U_i/N}{2\mu_i(1-U_i)} \tag{28}$$

when $L\to 1$. Observing that $GB(N)$ must have $\beta_i/N=\mu_i/\sum_{j=1}^N\mu_j$, $\forall i$, (as $L\to 1$) and that these terms equal the routing probabilities of a Bernoulli routing scheme, a term-by-term comparison between (28) and

$$R_{Bernoulli}^{So}=\sum_{i=1}^{N}\frac{\mu_i}{\sum_{j=1}^N\mu_j}\frac{1}{\mu_i(1-U_i)} \tag{29}$$

implies that $GB(N)\geq R_{Bernoulli}^{So}(N)/2$. As N grows, the terms related to the variance of the input process, i.e., U_i/N, approach zero and, as $L\to 1$, $U_i\to 1$, meaning that $GB(N)\to R_{Bernoulli}^{So}(N)/2$.

Proof of Theorem 3

Applying the standard analysis of $G/M/1$ queues, e.g., [7], it follows that

$$R^{\Gamma(N/\beta_i,N\lambda)/M/1}=\frac{1}{\mu_i(1-x)} \tag{30}$$

where x is the least positive solution of

$$z=\left(\frac{N\rho_i}{N\rho_i+1-z}\right)^{N/\beta_i}, \tag{31}$$

and $\rho_i=\lambda/\mu_i$, which we rewrite as

$$z\left(1+\beta_i\frac{1-z}{NU_i}\right)^{N/\beta_i}=1, \tag{32}$$

Assuming $a = (1-z)/U_i$, with a Maclaurin series expansion in β_i/N we obtain

$$\begin{aligned}\exp(-a)\left(1+\beta_i\frac{a}{N}\right)^{N/\beta_i} = \; & 1-\tfrac{1}{2}a^2\frac{\beta_i}{N} \\ & +\left(\tfrac{1}{3}a^3+\tfrac{1}{8}a^4\right)\frac{\beta_i^2}{N^2} \\ & -\left(\tfrac{1}{4}a^4+\tfrac{1}{6}a^5+\tfrac{1}{48}a^6\right)\frac{\beta_i^3}{N^3} \\ & +O(\beta_i^4/N^4)\end{aligned} \tag{33}$$

where the coefficient of $(\beta_i/N)^{-i}$, $i \geq 0$, alternates because $a > 0$. Observing that σ and σ^+ refer to truncations of the alternating series above, we must have $\sigma \leq x \leq \sigma^+$, which implies (10).

Proof of Lemma 1

The fact that $\sigma_i \leq U_i$ follows by the following facts. Let $f(z) = z\exp(-\frac{z}{U_i}) - \exp(-\frac{1}{U_i})$. We have:
i) $f(U_i) > 0$,
ii) $f'(z) = \exp(-\frac{z}{U_i})(1 - z/U_i) = 0$ if and only if $z = U_i$,
iii) $f''(z) = -\frac{1}{U_i}\exp(-\frac{z}{U_i})(2 + z/U_i) < 0$ $(z \geq 0)$, i.e., $f(z)$ is concave, and
iv) (8) has only two (positive) real roots (when $0 \leq U_i \leq 1$) where the largest one is $z = 1$.

Now, taking into account facts i)–iv), the statement can be proven by showing that $f(z)$ is positive when $z = U_i^2$. This simplifies to

$$h(U_i) = 2U_i \ln U_i - U_i^2 + 1 \geq 0. \tag{34}$$

Since

$$\frac{1}{2}\frac{\mathrm{d}h(U_i)}{\mathrm{d}U_i} = \ln U_i + 1 - U_i < 0, \forall U_i \in [0,1), \tag{35}$$

which easily follows by the change of variable $U_i = 1 - x_i$ and expanding the logarithm in Taylor series, and

$$\lim_{U_i \to 1} h(U_i) = 0, \tag{36}$$

we conclude that $h(U_i)$ must be strictly positive when $U_i \in [0,1)$.

Proof of Theorem 4

Within the $GB(N)$ bound (6), the response time of queue i is $R^{\Gamma(N/\beta_i, N\lambda)/M/1} = 1/(\mu_i(1-\sigma_i))$, where σ_i is the least positive root of

$$z\left(1+\frac{1-z}{N\rho_i}\right)^{N/\beta_i} = 1. \tag{37}$$

After a Maclaurin expansion in β_i/N (see (33)) and taking the first two expansion terms (yielding (8) and (9)), we must have $GB(N) - GB(\infty) > GB_N(\infty) - GB(N)$. The statement follows by observing that $GB(N) \leq R^{So}$.

Proof of Theorem 5

Given that the Hessian of the objective function is diagonal and the constraints are linear, to prove the convexity of (15), it suffices to show that $U_i/(1-\sigma_i)$ is convex in U_i. Let $g = g(U_i) = -\exp(-1/U_i)/U_i$. From the expressions of the derivatives of the Lambert W function [11], we obtain

$$\frac{\mathrm{d}}{\mathrm{d}g}W(g) = \frac{W(g)}{g(1+W(g))}, \tag{38}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}g^2}W(g) = -\frac{\exp(-2W(g))(g+2)}{(1+W(g))^3} \tag{39}$$

and substituting $g = W(g)\exp(W(g))$ in the latter, which follows by the definition of the W function, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}U_i}\frac{U_i}{1+U_iW(g)} = \frac{1}{(1+UW(g))(1+W(g))} \tag{40}$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}U_i^2}\frac{U_i}{1+U_iW(g)} = \frac{-W(g)}{(1+W(g))^3U_i^2} \tag{41}$$

which is strictly positive because $0 < -W(g) < 1$, for $0 < U < 1$.

Proof of Theorem 6

Let

$$f_1(\mathbf{U}) = \frac{1}{\lambda N} \sum_{i=1}^{N} \frac{1}{2} \frac{U_i}{1-U_i} \tag{42}$$

for $\mathbf{U} \in \mathbb{R}^N : \sum_{i=1}^{N} \frac{\mu_i}{\lambda} U_i = N,\ 0 \le U_i < 1,\ \forall i$ and

$$f_2(\mathbf{U}) = \frac{1}{\lambda N} \sum_{i=1}^{N} \frac{U_i}{1 + W(g(U_i))U_i} \tag{43}$$

for $g(U_i) = -\exp(-1/U_i)/U_i$ and $\mathbf{U} \in \mathbb{R}^N : \sum_{i=1}^{N} \frac{\mu_i}{\lambda} U_i = N,\ 0 \le U_i < 1,\ \forall i$. To prove the theorem, we show that $f_2(\mathbf{U}) \ge f_1(\mathbf{U})$, $\forall \mathbf{U}$, which is equivalent to show

$$1 + W(g(U_i))U_i \le 2(1 - U_i),\ \forall i, \tag{44}$$

Since $-W(g(U_i) \le U_i$ by Lemma 1, (44) is satisfied because it holds with equality when $U_i = 1$ and strictly when $U_i = 0$, and

$$\frac{\mathrm{d}}{\mathrm{d}U_i} \left[\frac{1}{U_i} - 2 - W(g(U_i)) \right] = -\frac{1}{U_i^2} - \frac{(1-U_i)W(g(U_i))}{U_i^2(1 + W(g(U_i)))} \tag{45}$$

is always negative.

Proof of Corollary 4

Monotonicity

From the expressions (38) and (39), we have

$$\frac{\mathrm{d}S(U)}{\mathrm{d}U} = \frac{UW(g)^2 + W(g) + W(g)U^2 + U}{U(1 + W(g))(1 - W(g))^2} > 0 \tag{46}$$

if and only if $UW(g)(W(g) + U) + W(g) + U > 0$. Lemma 1 implies that $-W(g) \le U_i$ and proves that it is increasing in U.

Limit as $U \to 1$

From the expressions (38) and (39), we have

$$\frac{\mathrm{d}A(U)}{\mathrm{d}U} = \frac{UW(g)^2 + W(g) + W(g)U^2 + U}{U(1 + W(g))(1 - W(g))^2} > 0 \tag{47}$$

and applying L'Hôpital's rule, we obtain

$$\begin{aligned} &\lim_{U \to 1} \frac{\mathrm{d}A(U)}{\mathrm{d}U} \\ &= \lim_{U \to 1} \frac{UW(g)^2 + W(g) + W(g)U^2 + U}{4(1 + W(g))} \\ &= \lim_{U \to 1} \frac{W(g)^2 + 2UW(g)W'(g) + W'(g) + 2UW(g) + W'(g)U^2 + 1}{4W'(g)} \end{aligned} \tag{48}$$

Since $\lim_{U \to 1} W(g) = \lim_{U \to 1} W'(g) = -1$ (see the proof of Corollary 17), the limit (47) is zero.

Limit as $U \to 0$

From (47), we obtain

$$\begin{aligned}
&\lim_{U\to 0}\frac{\mathrm{d}A(U)}{\mathrm{d}U}\\
&=\lim_{U\to 0} W(g)^2+\frac{W(g)}{U}+W(g)U+1\\
&=1+\lim_{U\to 0} W'(g)\\
&=1+\lim_{U\to 0}\frac{W(g)}{1+W(g)}\frac{1-U}{U^2}\\
&=1+\lim_{U\to 0}\frac{W(g)}{U^2}
\end{aligned} \tag{49}$$

The Lambert W function admits the following Maclaurin expansion [11]

$$W(g)=-\sum_{n\geq 1}\frac{n^{n-1}}{n!}(-g)^n \tag{50}$$

which is convergent $\forall g : |g| \leq 1/e$. The leading term of (50) is $\exp(-1/U)/U$, when $U \to 0$. Therefore, we have

$$\lim_{U\to 0}\frac{W(g)}{U^2}=\lim_{U\to 0}\frac{\exp(-1/U)}{U^3}=0. \tag{51}$$

References

[1] E. Altman, U. Ayesta, and B. Prabhu. Load balancing in processor sharing systems. In *ValueTools '08*, pages 1–10, ICST, Brussels, Belgium, Belgium, 2008. ICST.

[2] E. Altman, B. Gaujal, and A. Hordijk. *Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity*. Number 1829 in LNM. Springer-Verlag, 2003.

[3] Y. Arian and Y. Levy. Algorithms for generalized round robin routing. *Oper. Res. Lett.*, 12:313–319, 1992.

[4] U. Ayesta, O. Brun, and B. Prabhu. Price of anarchy in non-cooperative load balancing. In *INRIA tech. rep.*

[5] A. Bar-Noy, R. Bhatia, J. S. Naor, and B. Schieber. Minimizing service and operation costs of periodic scheduling. *Math. Oper. Res.*, 27(3):518–544, 2002.

[6] C. H. Bell and S. Stidham. Individual versus social optimization in the allocation of customers to alternative servers. *Management Science*, pages 29:83–839, 1983.

[7] U. N. Bhat. *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Birkhauser Verlag, 2008.

[8] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[9] H.-L. Chen, J. R. Marden, and A. Wierman. The effect of local scheduling in load balancing designs. *SIGMETRICS Perf. Eval. Rev.*, 36(2):110–112, 2008.

[10] M. B. Combé and O. J. Boxma. Optimization of static traffic allocation policies. *Theor. Comput. Sci.*, 125(1):17–43, 1994.

[11] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jerey, and D. E. Knuth. On the lambert w function. *Adv. Comput. Math*, pages 329–359, 1996.

[12] S. C. Dafermos and F. T. Sparrow. The traffic assignment problem for a general network. *J. Res. Nat. Bureau Standards,B*, 73(2):91–118, 1969.

[13] E. J. Friedman. Genericity and congestion control in selfish routing. In *43rd IEEE Conf. on Decision and Control*, pages 4667–4672, 2003.

[14] B. Gaujal, E. Hyon, and A. Jean-Marie. Optimal routing in two parallel queues with exponential service times. *Discrete Event Dynamic Systems*, 16(1):71–107, 2006.

[15] X. Guo, Y. Lu, and M. S. Squillante. Optimal probabilistic routing in distributed parallel queues. *SIGMETRICS Perf. Eval. Review*, 32(2):53–54, 2004.

[16] B. Hajek. The proof of a folk theorem on queueing delay with applications to routing in networks. *J. ACM*, 30:834–851, 1983.

[17] B. Hajek. Extremal splitting of point processes. *Math. Oper. Res.*, 10:543–556, 1986.

[18] M. Haviv and T. Roughgarden. The price of anarchy in an exponential multi-server. *Oper. Res. Lett.*, 35(4):421–426, 2007.

[19] A. Hordijk and D. van der Laan. Periodic routing to parallel queues and billiard sequences. *Mathematical Methods of Operations Research*, 59(2):173–192, 2004.

[20] H. Kameda and O. Pourtallier. Paradoxes in distributed decisions on optimal load balancing for networks of homogeneous computers. *J. ACM*, 49(3):407–433, 2002.

[21] F. Kelly. Loss networks. *Ann. Appl. Prob.*, 1:319–378, 1991.

[22] F. Kelly. Network routing. *Philosophical Transactions of the Royal Society A337*, pages 343–367, 1991.

[23] J. F. C. Kingman. Some inequalities for the queue gi/g/1. *Biometrika*, 49(3/4):315–324, 1962.

[24] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *STACS*, volume 1563 of *LNCS*, pages 404–413, January 1999.

[25] E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik. *Quantitative system performance: computer system analysis using queueing network models.* Prentice-Hall, Upper Saddle River, NJ, USA, 1984.

[26] J. S. Matthew and P. H. Daniel. *Stochastic Models in Operations Research, Vol. II: Stochastic Optimization.* Dover, December 2003.

[27] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory.* Cambridge University Press, New York, NY, USA, 2007.

[28] C. H. Papadimitriou. Algorithms, games, and the internet. In *ICALP: Proc. of the 28th Int. Colloquium on Automata, Languages and Programming*, pages 1–3, London, UK, 2001. Springer-Verlag.

[29] T. Roughgarden. The price of anarchy is independent of the network topology. In *J. of Computer and System Sciences*, pages 428–437, 2002.

[30] R. Subrata and A. Y. Zomaya. Game-theoretic approach for load balancing in computational grids. *IEEE Trans. Parallel Distrib. Syst.*, 19(1):66–76, 2008.

[31] A. Wächter and L. T. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

[32] J. Walrand. *An Introduction to Queueing Networks.* Prentice-Hall, 1988.

[33] W. Whitt. *Stochastic Process Limits.* Springer, New York, 2002.

[34] T. Wu and D. Starobinski. On the price of anarchy in unbounded delay networks. In *GameNets*, page 13, New York, NY, USA, 2006. ACM.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399